An Integrated Approach of Mental Workload Assessment

Marc Schneider and Barbara Deml

Abstract

There are a lot of methods and instruments trying to assess mental workload reliable and independent from specific use cases or situations. Each method has its own strengths and weaknesses depending on the situation and the on individual person. Attempting to improve each approach separately would indeed allow a further context of use, but do not lead to a global reliability growth. Within this work a new approach is suggested by combining various ocular parameters adapting the individual person. The results of this study show clearly the advantages in mental workload assessment of an individual adapted set of variables in contrast to one variable trying to fit a collectivity of people. This new preparation considering six physiological ocular variables in two different contexts of use can be one step further to a more reliable and global way to assess mental workload.

M. Schneider (✉)
Institut für Arbeitswissenschaft und Betriebsorganisation, Karlsruher Institut für Technologie, Karlsruhe, Germany
e-mail: marc.schneider3@kit.edu

B. Deml
ifab-Institute of Human and Industrial Engineering, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
e-mail: barbara.deml@kit.edu

1   Introduction

The reliable assessment of mental workload is both a fundamental and challenging requirement for a user-centered, adaptive workplace design in human-machine-interaction. This is particularly true against the context of older workforce and the increasing complexity of systems. There is already a diversified portfolio of approaches promising the best possible assessment of mental workload which rely on different procedures of subjective, performance as well as physiological measurements. Each of these approaches has certain strengths and weaknesses concerning its reliability [21]. This variability can be assigned to different contexts of use or various experimental settings.

The mental workload itself is a very complex phenomenon which is widely analyzed theoretically as well as empirically for many years until today [13]. There are lots of models and constructs trying to describe cognitive processing and mental workload which are subject to constant expansion, upgrades or modifications [14, 17, 29]. One approach to describe mental workload is based on the resource theory. A human has an individual amount of cognitive capacities to handle different tasks with a specific demand to these resources. On the one hand the demand varies depending on several factors like the time on task, the modalities or the level of complexity and on the other hand the capacity is individually different. Each individual has different resources depending not only on his or her constitution, but also on level of training on the specific task. If the demands exceed the capacity of resources available for a specific task it comes to an overload as well as the other way round too low demands can lead to underload [39, 40]. This approach gives one example of how mental workload and its formation can be described. Besides the above mentioned and often in studies for workload assessment named models for cognitive processing and workload there is another essential theoretical construct considered in this work which finds relatively little application in human factors. In high stress or high workload situations identified by subjective, objective or performance (secondary tasks) measures, the individual workload depends on the particular coping strategy of the human [12, 27, 30]. In general the individual coping can be divided into two modes, the active and the passive coping [30]. The active coping contains conscious adaptions of behavior like changing the amount of effort spent or the prioritized task in a dual-task configuration. Changes in processing strategies or the ignorance of certain task components can also be assigned to the active coping [27]. Passive coping includes the unconscious adaption such as the regulation of the information flow and processing. All in the entire individual coping considers both external (e.g. time pressure, negative feedback) and personality (e.g. experiences, constitution, level of skill) factors [27]. The type of coping depends among others on the subjective perceived level of task-induced demand. Robert and Hockey [30] introduced a two level system. Up to an individual depending amount of demand the human is able to cope with workload passively whereas demand higher than this point can only be regulated by active coping [30]. Nevertheless it is not possible to predict the coping strategies from the task-induced demand alone due to the individual perception and handling with the situation. The individual coping is time-dependent and it can also change

over short intervals of time making a reliable prediction of workload additionally difficult [30].

In this work six correlates (Blink Rate, Blink Duration, PERCLOS, Pupil Dilations, Fixation Duration and Nearest Neighbor Index) for mental workload assessment, that are based on eye-tracking data compared in two different scenarios with each three levels of difficulty. Within the first setting a flight simulator task is regarded, whereby the autopilot is defect, so that both the altitude and the heading are to be kept manually. The stress level is varied by the presence of a secondary task: while there was no secondary task in one condition, the participants had to do either an acoustic or a visual 1-back secondary task in the further two conditions. For the 1-back task the subjects have to repeat verbally the last number each time a new one is presented in the both mentioned modalities [19]. Within the second setting a mental rotation task is studied in which two geometries have to be checked for equality. Here the stress level is varied by the number of contour points. As both experimental settings require different cognitive efforts, the workload assessment can be tested in two different contexts of use. The flight simulator task is characterized by a dynamic situation with simultaneous tasks whereas the mental rotation is a sequential and one dimensional setting. Besides procedures regarding the pupil an eye behavior like the pupil dilations and the blink rate the study considers also the gaze strategy operationalized by the nearest neighbor index. The NASA-TLX questionnaire is used to evaluate the subjective perception of the effort needed in the tasks.

The reminder of the article is structured as follows: At first each method is explained theoretically before introducing the experimental setting. Afterwards the results were described before closing with a discussion.

2   Workload Measurement Methods

Commonly three different categories of workload measurement techniques are distinguished: performance measures (e.g. dual-task paradigm) [40], subjective rating scales [31], and psychophysiological measures [21]. Depending on the research question, different measurement instruments may be used for each of the three categories. Within this section, only the last two categories shall be regarded, whereby as self-assessment technique NASA Task Load Index (NASA-TLX) is used [16]. This questionnaire is rather widespread as it is economical, clearly structured and easy to understand [31]. NASA-TLX consists of six items (Mental Demand, Physical Demand, Temporal Demand, Performance, Effort and Frustration Level), whereas the item "physical demand" is not relevant for our study and thus, it was not regarded here.

Psychophysiological measures (e.g. heart rate, skin conductance and electroencephalography) examine in an objective and continual way what effect mental workload has on an operators' body. Despite of these advantages the methods are quite often more general indicators of stress and thus, it is of particular importance to assure that they are relevant for the research question at hand. In our study only

ocular parameters shall be regarded because of the close linkage of the eyes to the information processing regions of the human [37]. Due to the mentioned problem of sensitivity, we decided to analyze various ocular parameters that differ in terms of temporal sensitivity—the blinking behavior Sect. 2.1, the pupil dilations Sect. 2.3, the fixations duration Sect. 2.4, as well as gaze strategies Sect. 2.5. While, for instance, pupil dilations distinguish by a high temporal resolution, other variables, such as gaze behavior have to be aggregated over time.

Due to the time dependency of the individual coping strategies [30], the following explained parameters are calculated with different resolutions having diverse sensitivities to fluctuations over time. In the case of NNI and pupil dilations there are the mean, median, maximum value and minimum value considered in this study. Especially for the pupil dilations it is additionally possible to calculate the mean of the first values which is equivalent to the first 30 s of the task. On this way it might be possible to detect individual changes with respect to coping strategies.

2.1   Blink Rate

Commonly, an eye blink takes about 70–100 ms [22] and the *blink rate* describes the number of eye closures in a pre-defined period of time. Usually, this time period is set to 1 min, but in the case of very short or variable tasks an interval of 30 s may be analyzed too [41]. The blink rate is often used to measure mental workload. Wilson [41], for example, showed that there is a significant negative relationship between the blink rate and the increasing difficulty of certain temporal segments of a flight in a flight simulator. However, for example, Cardona and Quevedo [7] could not find any significant correlation between the blink rates of car drivers and their performance at different complexity levels. This may also be due to the fact that blink rate is not only influenced by the mental workload, but it is also altered by other user states, such as the degree of sleepiness [38]. There is a seamless continuation between a blink rate of a non-sleepy user and that of a sleepy user, whereby a blink rate of longer than 500 ms may be referred to as micro-sleep [33]. Thus, it is reasonable to combine the blink rate with other indicators of mental workload in order to be more confident that the observed differences are really due to the workload of the user and that they are not caused by other factors.

2.2   Blink Duration and PERCLOS

As shortly mentioned above the *blink duration* can take between 70 and 500 ms. Benedetto et al. [4] suggest that the blink duration is even more sensitive to the cognitive workload than the blink rate. They tested both parameters in a lane change task with two conditions, control condition and dual task condition. Their results showed that the blink duration was shorter for more difficult tasks, whereby no significant effect could have been found for the parameter blink rate [4]. Martins and Carvalho [26], too, analyzed the blink characteristics of users in the context of mental workload and fatigue. In their study the parameter *PERCLOS* (percentage of

eye closure) was derived, which is a rather well established parameter to predict fatigue and which is defined as the accumulated blink duration on a pre-defined time interval [1]. Brookhuis and de Waard [6] also assessed PERCLOS and they have shown that this parameter is significantly correlated to the performance of car drivers in a driving simulation. Due to its similarity to the blink duration and the promising results of former studies, the parameter PERCLOS is also taken into account for mental workload assessment in this study.

2.3 Pupil Dilations

The pupil of the human eye may be compared to an optic lens, which changes its size in order to regulate the incidence of light into the eye. The pupil dilations are controlled by two antagonistic muscles surrounding the pupil: the Sphincter Pupillae, which scales down the pupil size para-sympathetically, and the Dilator Pupillae, which scales up the pupil size sympathetically [3]. Some changes in pupil size, the so called light reflexes, adapt the pupil to the light conditions. In bright environments the pupil is small and contracted whereas in dark conditions the pupil is wide and relaxed. This adaption is characterized by smooth and lower frequent fluctuations in pupil size. Beside the lighting conditions, the pupil size is also sensitive to the viewing distance [3]. Furthermore the psycho-sensory reflex is to be mentioned [20]. This reflex is triggered by neurocognitive activities over different muscles depending paths, indicating that the fluctuation in pupil size is also related to mental workload [36]. A huge amount of studies under various conditions have been carried out in order to examine the influence of cognitive workload on the pupil reflex [18]. A major challenge of these studies was to extract the psycho-sensory reflex from the other changes in pupil size like the light reflex. Marshall [22] introduced the Index of Cognitive Activity (ICA), which succeeded the separation of the psycho-sensory reflexes. In the presence of psycho-sensory reflexes the signal, created from the continuous recording of pupil size, is characterized by abrupt discontinuities. In the presence of effortful cognitive processing, the pupil responds rapidly with a reflex reaction [23]. Marshall [22] uses the wavelet analysis which allows analyzing the frequency of a signal without losing the reference of time. There are several recent studies among others in the driver and the aviation context which show explicit the significant relation between the ICA and subjective as well as objective reference measurements [24, 25, 35]. Due to this wide spread positive responses to the promising approach of Marshall [22] the analysis of pupil dilation used in this work is also based on the wavelet analysis to separate the different pupil reflexes. The analysis is built on the Daubechies db 4 wavelet families due to the 60 Hz eye-tracking system used in the experiments [22]. Each type of wavelet families is suitable for a specifically defined frequency range of the output signal [11]. The signal of the pupil diameter orthogonally wavelet transformed and disassembled in a high and a low pass range. In this first step the intervals with high frequencies (detail) are separated from low frequencies intervals (approximate) without losing the reference of time. This transformation is a so called level 1 wavelet transformation [28]. In further levels

of transformation the approximate interval can be further divided step by step, but that is not necessary for the pupil analysis. After the transformation, a threshold is used to extract the typical irregular and high frequent changes in pupil diameter caused by cognitive arousal. The thresholding can be done by several types from soft over mini-max to hard thresholds. Each of the mentioned thresholds has its own strengths and weaknesses depending on analyzed signal structure [10, 11, 22, 32]. In this study the hard threshold is used with respect to the chosen signal decomposition technique. After setting a threshold the point of the signal below this border were set to zero, so that only the values corresponding to the criteria of a psycho-sensory reflex remain [22]. The number of reflexes per second is set in relation to a maximum number of 30 per second in order to become scaled value of dilations per second [8].

2.4   Fixation Duration

For human beings sharp seeing with a high resolution is only possible at a relatively small part at the center of the retina, which is known as Fovea Centralis. Due to this restriction, eye movements are needed, which ensure that the retinal image of the interesting object is mapped as accurately as passible on the retina. The alignment of the fovea, the "gaze" on an object is called fixational eye movement. It could have been shown that under high stress the *fixation duration* is directly related to the time, which is needed to either gather information or to understand a certain problem at hand [2]. In other words, if the mean fixation duration is short, a higher amount of time may be dedicated to searching behavior. In this study, too, the fixation duration is considered, whereby the mean value of all fixations in a defined period of time (30 s) is regarded.

2.5   Nearest Neighbor Index (NNI)

The *Nearest Neighbor Index* (NNI) is a spatial statistics algorithm, which in the context of eye-tracking expresses the proximity of each fixation relative to all other surrounding fixations. In other words, it assesses the randomness of fixation patterns. Primarily, this index was used to characterize the natural distribution of plant or animal populations, however, it can also be applied to gaze distribution patterns in order to identify gaze scanning strategies [9]. Thereby a ratio between the average of the observed minimum distances between fixations and the mean distance, which would to be expected for a random distribution of fixations, is derived. In consequence, for a random distribution a ratio that is equal to 1 is to be observed. For values less than 1 the mean of minimum observed distances is smaller than the random one, which indicates a clustering of fixations and which stands for so called informational gaze patterns. For values larger than 1 a regular, non-random gaze behavior is given [9]. According to Di Nocera et al. [9] the value range for NNI lies between 0, which corresponds to maximum clustering, and 2.1491, which accounts for a strictly regular hexagonal pattern. It is assumed

that under high stress conditions the gaze behavior does not follow a clear strategy and that it is characterized by searching for information randomly. In consequence, in these situations NNI should account for 1. There is a lot of empirical evidence that NNI reveals a significant correlation with both subjective as well as objective performance parameters of cognitive workload 15].

When applying NNI in a certain study, it is necessary to define the reference area containing all fixations of interest—otherwise neither the random distribution nor the actual distribution of fixation patterns may be derived. According to Di Nocera et al. [9] it is only required that the area is large enough to include all fixations. Within their work a rectangle, set by the lowest and highest values of each x and y coordinates of fixations, is used. This may work for homogeneous conditions without significant outliers, but it causes problems in more heterogeneous conditions, when there are also fixations outside the field of activity. If there is only one fixation, which is far away from the others, the rectangle will be enlarged and a different distribution, with respect to the reference area, will be provided. Against this context it may be reasonable to use a dynamic elliptical area based on a chi-square distribution. Within the work of Schubert and Kirchner [34] this approach was developed successfully evaluated against other statistical distributions in the field of postural analysis; it contained more than 95 % of all points and it adapted very well to the distribution of these points. Due to this dynamic adaption, which is able to cope with heterogeneous conditions, this approach is used to calculate the reference area for the NNI in this work.

3 Experimental Setting

In the experiment 48 subjects (32 males, 16 females) with a mean age of 22.44 years ($SD = 3.26$) took part; most of them were students of our university. In the context of an eye-tracking study it is also of interest to mention that 19 participants (13 glasses, 6 contacts lenses) had a corrected vision. The eye-tracking data were recorded by the 60 Hz binocular DIKABLIS Professional system of the Ergoneers GmbH.

The experiment consisted of two tasks, whereby each task revealed three stress levels. A within-subjects design was chosen and all participants had to pass all six conditions in a randomized order to avoid order effects. After each of the six conditions the participants had to complete NASA-TLX and further on the eye-tracking device was re-calibrated in order to ensure the accuracy of the recordings.

The first task consists of a flight procedure in a flight simulator. The autopilot is defect so that the test person has to control the plane manually. The main task is to keep the heading and the altitude on the target given by the flight director (Alt: 5000 ft.; HDG: 360°) Fig. 1.

The subject controls the plane via stick like in a real setting. On a screen in front of the person the primary flight display with all relevant information is shown. It is depicted greater than normal to focus the concentration on the necessary

Fig. 1 Experimental setting flight task
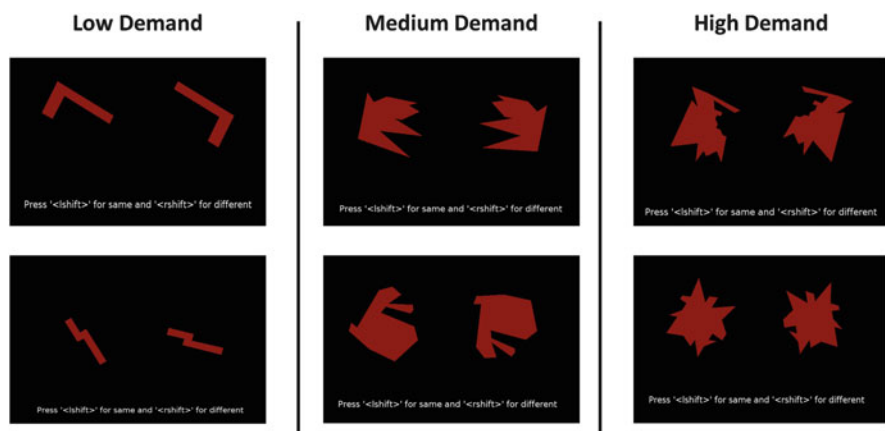
Fig. 2 Experimental setting mental rotation

parameters. Three stress levels have been operationalized by a secondary task: while the additional task was absent in a baseline condition, within two further conditions the participants had to complete a subsidiary task with a cycle time of 3 s, which was either announced acoustically or presented visually on a tablet PC left to the primary flight display Fig. 2. As secondary task a so called 1-back task was chosen. This is a rather widespread type of secondary task, which exists in various versions. Here, the participants had to reproduce verbally numbers on range from zero to ten presented every 3 s like mentioned above. Every time a new number was presented they had to represent the previous one. As an indicator for mental workload the number of mistakes was assessed [40].

The second task consists of a mental rotation in which two geometries had to be checked for equality. Here the three levels of stress (low, medium, high demanding) were defined by the number contour points of the figures Fig. 2.

The mental rotation test is developed in the experimental environment PEBL (Psychology Experimental Building Language). The geometries were shown in

randomized orientations and are either mirrored or equal [5]. The subject had to decide about the equality in between 3 s and if he or she was too slow the test jumped to the next pair of geometries. After every decision the participant became feedback (right/wrong) to keep the motivation up over the whole time. Every subject had to do 128 pairs of figures in each stress level with a fixed number of equal pairs in a randomized order. The program collects a set of information like the reaction times and the amount of mistakes, as well as the number of trails in which the participants were either too slow or did not answer.

4 Data Analysis and Results

The following data analysis is divided in three parts which constitute useful to each other. Starting with the subjective and the objective data across the physiological data and ending with a detailed analysis of subject specific data.

In a first step all the recording were reviewed to check them for any interferences or problems concerning the eye detection. There were six subjects with either defective recording from the field camera or defective recordings from the eye cameras, which could not be cleared by the recalibration. Therefore these six (Participants number 18, 24, 27, 33, 37 and 45) datasets had to be excluded from the further examination. Before starting the examination of the remaining datasets all the variables were analyzed using the Kolmogorov-Smirnov-Test for normal distribution which can also be adopted by the central limit theorem due to the size of the sample (~50). Each variable needed for an analysis of variance (ANOVA) was tested for homogeneity of variances by the Levene-Test before.

4.1 NASA-TLX and Performance Parameters

In a first step the subjective data (NASA-TLX) as well as the performance parameters are analyzed, which are used for a manipulation check of the physiological data. The performance parameters consist of the number of mistakes in the secondary tasks (flight task) and the reaction times as well as the number of mistakes in the comparisons of the geometries (mental rotation). Like mentioned before both test parts consist of three stress levels. At first it is necessary to know if the parameters are able to reproduce the diverse task-induced demands leading to different cognitive effort.

4.1.1 Flight Task

For the flight task there were only significant differences between the NASA-TLX values, $F(2, 94) = 59.16$, $p < .05$. The Bonferroni post-hoc test showed that the first setting without secondary task differs significant form both conditions with secondary task. Neither NASA-TLX nor the number of mistakes in secondary task induced significant differences regarding acoustic and visual condition. The effect described by these two parameters can be related to resource theory by Wickens et al. [40],

indicating that two visual tasks have a high inference and so a potential higher demand, but in this case there is often the effect of overlaying primary and secondary task. The task-induced various demands are hard to identify in the individual workload [40]. The subjective and objective data correlate significantly ($r = .40$, $p < .05$).

4.1.2 Mental Rotation

For the mental rotation test, an ANOVA showed a significant main effect between the three stress levels concerning NASA-TLX values, $F(2, 94) = 22.71$, $p < .05$. The pairwise comparison with Bonferroni revealed that NASA-TLX ratings of the most difficult level were significantly higher than of both other two situations. Also the number of mistakes in the mental rotation task differed between the stress levels, $F(21.76, 82.82) = 33.65$, $p < .05$. A Bonferroni post-hoc test indicated a significant higher number of mistakes in high demanding than in the medium demanding situation, whereas the low and the medium demanding situation differed not significantly. There are deviations between the subjective and the objective data. The correlation analysis confirmed this impression ($r = .07$, *n.s.*). The drift between NASA-TLX and number of mistakes can be attributed essentially to two possible reasons. On the one hand the usage of the NASA-TLX with untrained people can, despite the simplicity, cause distortion and on the other hand NASA-TLX may not be appropriate to map such a granular differentiation.

All in all it can be summed up that the subjective perception of mental workload operationalized by the NASA-TLX doesn't fit the objective performance parameters perfectly. It may be noted that the significant deviations in the NASA-TLX are accompanied by great jumps in the number of mistakes. The NASA-TLX scores moves around ten on the scale from 0 (low) to 20 (high) [31], indicating a medium task-induced demand in both conditions.

4.2   Physiological Data

Now the parameters tracked by the head mounted eye-tracking system, explained in the part methods, move into the focus Sect. 2. For this step of analysis the common way of regarding the mean value of the physiological parameters for every condition of both tasks is performed, before having a closer look at the individuals in the next chapter.

At first the various parameters are analyzed task-divided concerning their sensitivity to the diverse stress levels with an ANOVA, before comparing them to the subjective and performance parameters.

For the flight task there weren't any significant differences between the stress levels. As it can be seen in illustration 1.3 the certain parameters follow a remarkable and increasing trend like it was in the subjective or objective data, but these alterations lie on a very small range, so that there are no significant differences to be mentioned Fig. 3.
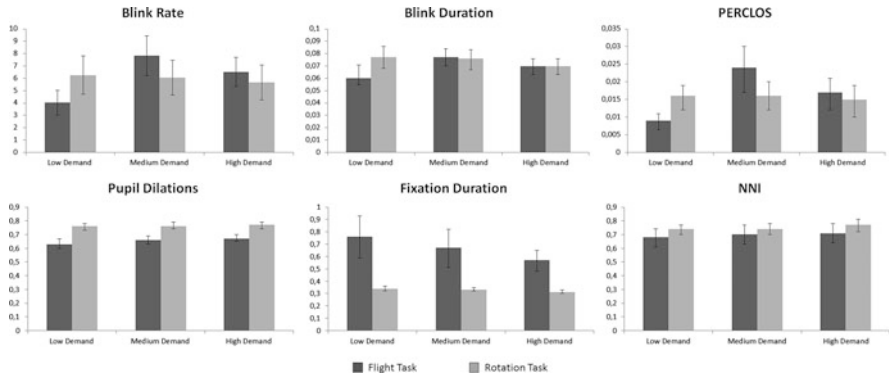
Fig. 3 Means of all physiological parameters for flight and rotation task

Like with the flight task there is none of the six parameters that indicate significant differences between the stress levels of the mental rotation task. Illustration 1.3 clarifies that by showing all six parameters Fig. 3. The majority of the parameters are stable on a specific level, so that there is no possible differentiation between the certain levels. In this task none of the parameters is neither able to detect differences between the levels of difficulty nor able to describe subjective or objective perceived workload.

After analyzing the sensitivity for the deviation in difficulty, the property of each parameter to predict the subjective sensation of mental workload as well as the objective parameters is considered. Therefore a correlation analysis is used. For the mental rotation task there are no significant correlation between the physiological measures and the subjective or objective data. The correlation coefficients are all close to zero indicating independence excepting a slightly positive relation between the number of pupil dilations and the NASA-TLX. For the flight task there are significant correlations. The NNI correlates significant with the NASA-TLX ($r = .27$, $p < .01$) as well as the number of mistakes in the secondary task ($r = .27$, $p < .01$). There is also a significant correlation between pupil dilations and NASA-TLX ($r = .23$, $p < .01$).

In summary it can be said that none of the used physiological eye-tracking measures is able to have a reliable mental workload assessment for both types of tasks in this sample of 48 subjects. In the flight task the NNI and the pupil dilations tend to fit well to the subjective data whereat the explained variance is still relatively low. There are mainly two possible reasons for the missing reliability of the physiological parameters. On the one hand a task dependence of the parameters can be postulated due to the differences between the tasks, and on the other hand the mentioned individual coping strategies may influence the reliability that much. These results motivate the following detail examination of the individual subjects concerning the coping strategies.
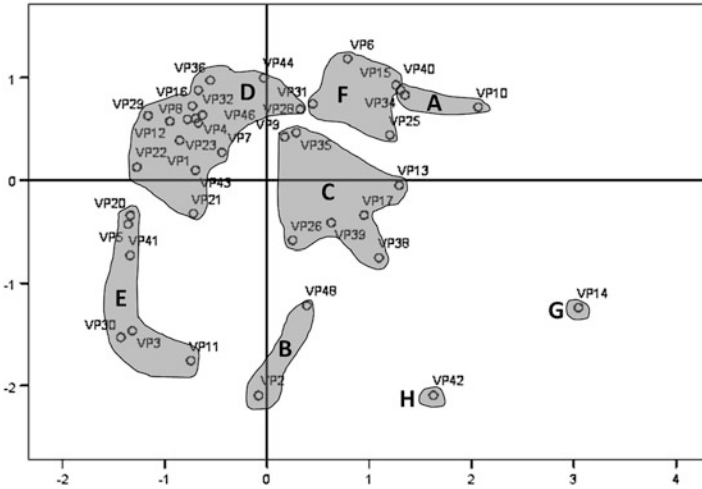
4.3   Detail Examination of Subjects

Based on the described theory of individual strategies for coping with high work-
load in the introduction (Sect. 1) and the results of the examination of the entire
sample (Sect. 4.2), it is necessary to have a closer look at the individuals. Therefore
the following analysis is divided into three steps. In the first part a cluster analysis
was used trying to find groups of subjects with similar characteristics in terms of
coping strategies. Afterwards each subject was analyzed separately over both test
parts with each three stress levels concerning the relation between the physiological
parameters, including the different temporal resolutions (Sect. 2), and the subjec-
tive as well as performance parameters. At least these two steps were considered
together looking for dependencies and conclusions.

4.3.1 Results of the Clustering Analysis

When all physiological parameters with all possible resolutions were considered,
there were 13 parameters per subject, which could be taken into account. These
13 parameters for both tasks with each three stress levels were considered for a
correlation between the subjects. The result was a triangle matrix with the
correlations between all subjects, which was the input for Multidimensional Scaling
(MDS). The MDS calculated the Euclidian Distance based on the correlation
coefficients between all subjects and plotted them into a two dimensional coordi-
nate system. A MDS of all 42 participants indicated that there were two outliers
being spatially wide separated from the remaining sample. A closer look at the
questionnaires of both subjects showed that they are real outliers due to their
NASA-TLX scores weren't higher than 1 in any tasks or stress levels as well as
their number of mistakes, which were nearly zero. The results of another MDS
carried out after ignoring these two participants indicated that there were certain
groups of subjects with similar characteristics Fig. 4. To specify this visual impres-
sion a k-means analysis was accomplished. For this type of clustering algorithm a
desired number of clusters had to be defined before running the analysis. The
k-means analysis is then followed by an ANOVA checking the found clusters for
differences. After trying all possible numbers of clusters (1–40), the best fitting
solution contains eight clusters. These eight clusters (A to H) significantly differed
from each other, so that this solution could be regarded as robust Fig. 4.

If both tasks (flight task & rotation task) are regarded separately, the k-means
analysis prevents also eight clusters which are significantly different from each
other. The main positions of the subjects in the coordinate system and their cluster
allocation are consistent with the overall picture Fig. 4. Since the variations in the
precise mapping are rather low, the hypothesis of task dependency of the
parameters (Sect. 4.2) can be discarded.

For the last step of this cluster analysis, the naming of the axes, it is useful to
consider the following individual examination of all variables, so that the naming is
placed below this step.

Fig. 4 Results of MDS with k-means clusters

4.3.2 Results of the Detail Examination of all Parameters

Another step further to a reliable naming of the axes concerning the extracted clusters is to have a closer look at each individual. Each of the 13 variations of physiological parameters was correlated to the number of mistakes as well as the NASA-TLX scores for every subject. Due to the independence of the task, all six conditions of both tasks were considered, so that there are six values for every parameter. A first look at the correlation coefficients showed that there are high significant relations between the number of mistakes and different parameters for each subject. Considering the clusters found by the k-means analysis, it turned out that the highly significant predictors for the number of mistakes and the NASA-TLX distinguish between the clusters, whereas within the clusters the parameters are homogeneous. For example all subjects in cluster A can be significantly described by mean of the pupil dilations, the min values of the NNI and also by the Blink Rate as well as the PERCLOS. In contrast cluster B can be described by the first values of the pupil dilations and the maximum value of the NNI. If each cluster is analyzed with the best fitting parameters, the mean of the correlations between these physiological parameters and the performance parameters for the whole sample is highly significant and shown in Table 1.

The differences in the sample indicated by the clusters can be confirmed by the detail examination of physiological parameters for each subject.

4.3.3 Development of Coordinate System

Taking into account both, the results of the cluster analysis and the differences found concerning the best predicting physiological parameters, a coordinate system can be developed, which refers to the theory of individual coping strategies Sect. 1.

Table 1 Means of correlation coefficients between the physiological data and the performance data

	Blink rate	Blink duration	PERCLOS	Pupil dilations	Fixation duration	NNI
Number of mistakes	$r = -.79$, $p < .05$	$r = -.80$, $p < .05$	$r = -.78$, $p < .05$	$r = .84$, $p < .05$	$r = -.81$, $p < .05$	$r = .79$, $p < .05$
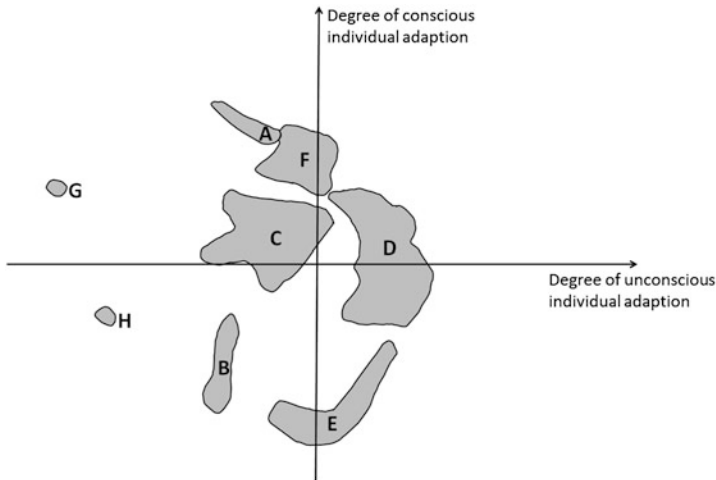
Fig. 5 Clusters after rotation of axes

After rotating the coordinate system around 45°, which is a common method in cluster analysis, an explicit assignment can be found (Fig. 5).

The X-Axis can be named with "Degree of unconscious adaption" according to the unconscious adaption or coping with high workload, whereas the Y-Axis shows the degree of conscious adaption of the individual to high task-induced demands Sect. 1. In terms of the pupil dilations it is remarkable that the best predicting resolution changes along the X-Axis. For example cluster D can be better described by the first values of the pupil dilations, whereas clusters B, G and H can be best described by the mean value. For group D the mean of the first values of the pupil dilations are significantly higher than the following values, $t(226) = -2.0, p < .05$. The indicated workload by the pupil dilations decreases over the process of one specific stress level, whereas in groups, which can be better predicted by the mean value, the pupil dilations stay stable over the time. This effect indicates that the subjects, best predicted by the first values, develop unconscious individual adopting strategies. The NNI describes the gaze behavior or strategy of individuals. For example group A can be best described by the minimum values, whereas group B can be best described by the maximum value and for group C the mean value fits best. Regarding the NNI for these three groups of individuals over the time of one

Table 2 Changing parameters concerning the defined axes

	High	Low
Degree of unconscious individual adaption (X-Axis)	Pupil dilations decrease over task time Blink Duration as a good predictor	Pupil dilations stay stable over time Blink Duration not a significant
Degree of conscious individual adaption (Y-Axis)	Changing individual gaze behavior over task time Blink Rate as a good predictor Low individual effort (NASA-TLX)	No changes in gaze behavior over task time Blink Rate as moderate predictor

stress level, it becomes clear the NNI changes over the time for the subjects with maximum or minimum values as best predictors. In case of the minimum values the curves look like bath tubs, whereas for maximum values the bath tubs are inverted and the mean values stay stable. In a further step the data stream is cut in three intervals with borders around the areas of the curves with the steepest decreases or increases. The ANOVA shows significant differences between the levels of the three time intervals for the minimum value group, $F(2, 210) = 2.84$, $p < .05$. The subjects for who the mean is the best fitting value don't show this effect. So there is a significant change in the individual gaze behavior or informational strategy over the time for those who can be best described by the minimum values of the NNI.

It can be seen that the theory of conscious und unconscious coping strategies can be confirmed by these parameters. Also the other recorded parameters change along the new defined axes, which are summarized by the following Table 2.

All in all it can be said, that depending on the individual coping strategies the sample can be divided in eight significant clusters. After a 45° axes rotation in combination with the individual examination of various variants of all physiological parameters there are significant differences between these clusters concerning the different adaption types described in Sect. 1. This adaption of physiological parameters can obtain very good task overreaching results concerning the reliability of workload assessment.

5   Discussion

This study shows the application of a portfolio of promising physiological eye-tracking measures for mental workload assessment. Therefore two tasks with each three different stress levels were chosen to analyze the parameters in different cases of use. The analysis clarifies that the individual coping strategies have a significant impact on the reliability of mental workload assessment. A clustering of the sample accompanied by an individual examination of all the introduced methods and instruments provide a substantial increase in reliability of mental workload assessment. There is for each subject at least one parameter which significantly represents the mental workload operationalized by the number of

mistakes as well as the NASA-TLX. The combination is very variable due to the different clusters in the defined coordinate system (conscious and unconscious individual adaption to workload). It is possible to assess mental workload reliable and additionally independent from the task. An upcoming problem with this approach is the need of an objective measure or parameter for workload like the number of mistakes in a secondary task for the differentiation in the physiological parameters, which is needed together with the clustering for the integrated workload assessment. In most practical experimental setting this is hard to realize. A required further step in the development of such an approach for reliable mental workload assessment is the generalization for all, laboratory and field, settings. Therefore our institute is currently in the intensive development of software with which a fast calibration of any eye-tracking system is introduced. The software calculates automatically all named physiological measures in all variations and adapts them to the individual subjects coping strategies based on objective data and the cluster analysis.

References

1. Abe T, Nonomura T, Komada Y, Asoaka S, Sasai T, Ueno A, Inoue Y (2011) Detecting deteriorated vigilance using percentage of eyelid closure time during behavioural maintenance of wakefulness tests. Int J Psychophysiol 82:269–274
2. Backs R, Walrath L (1992) Eye movement and pupillary response indices of mental workload during visual search of symbolic displays. Appl Ergon 23:243–254
3. Beatty J, Lucero-Wagoner B (2000) The Pupillary System. In: Cacioppo J, Tassinary L, Berntson G (eds) Handbook of psychophysiology. University Press, Cambridge, pp 142–162
4. Benedetto S, Pedrotty M, Minin L, Baccino T, Re A, Montanari R (2011) Driver workload and eye blink duration. Transp Res F 14:199–208
5. Berteau-Pavy D, Raber J, Piper B (2011) Contributions of age, but not sex, to mental rotation performance in a community sample. PEBL Tech Rep Ser 11(2)
6. Brookhuis K, de Waard D (2010) Monitoring drivers' mental workload in driving simulators using physiological measures. Accid Anal Prev 42:898–903
7. Cardona G, Quevedo N (2014) Blinking and driving: the influence of saccades and cognitive workload. Curr Eye Res 39:239–244
8. Demberg V (2013) Pupillometry: the index of cognitive activity in a dual-task study. In: Knauff M, Pauen M, Sebanz N, Wachsmuth I (eds) Proceedings of the 35th annual conference of the cognitive science society. Austin TX, Berlin
9. Di Nocera F, Camilli F, Terenzi M (2007) A random glance on the flight deck: pilots' scanning strategies and the real-time assessment of mental workload. J Cogn Eng Decis Making 1 (3):271–285
10. Donoho D (1995) De-noising by soft-thresholding. IEEE Trans Inf Theory 41:613–627
11. Donoho D, Johnstone I (1995) Adapting to unknown smoothness via wavelet shrinkage. J Am Stat Assoc 90:1200–1224
12. Gagnon J, Durantin G, Vachon F, Causse M, Tremblay S, Dehais F (2012) Anticipating human error before it happens: towards a psychophysiological modal for online prediction of mental workload. Proceedings of the human factors and ergonomics society 2012, pp 1–11

13. Gawron V (2008) Human performance, workload, and situational awareness measures handbook. CRC Press, Boca Raton, FL
14. Gopher D, Donchin E (1986) Workload – an examination of the concept. In: Schieber F (ed) Handbook of perception and human performance. Wiley, New York, NY, pp 1–49
15. Halverson T, Estepp J, Christensen J, Monnin J (2012) Classifying workload with eye movements in a complex task. Proceedings of the human factors an ergonomics society annual meeting, 58, pp 168–172
16. Hart S, Staveland L (1988) Development of the NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Hancock P, Meshkanti N (eds) Human mental workload. North-Holland, Amsterdam, pp 139–183
17. Kahneman D (1973) Attention and effort. Prentence-Hall, Englewood Cliffs, NJ
18. Kahneman D, Beatty J (1967) Pupillary responses in a pitch-discrimination task. Percept Psychophys 2:101–105
19. Kane M, Conway A, Miura T, Colfesh G (2007) Working memory, attention control, and the n-back task: a question of construct validity. J Exp Psychol 33:615–622
20. Loewenfeld I (1993) The Pupil: anatomy, physiology and clinical applications, 1. Iowa State University Press, Ames, IA
21. Manzey D (1998) Psychophysiologie mentaler Beanspruchung. In: Rösler F (ed) Enzyklopädie der Psychologie. Hofgrefe, Göttingen, pp 799–864
22. Marshall S (2000) Patentnumber. 6,090,051. USA
23. Marshall S (2002) The index of cognitive activity: measuring cognitive workload. IEEE Human Factors Meeting, 7, Scottsdale
24. Marshall S (2007) Indentifying cognitive state from eye metrics. Aviat Space Environ Med 78:165–175
25. Marshall S, Pleydell-Pearce C, Dickson B (2003) Integrating psychophysiological measures of cognitive workload and eye-movements to detect strategy shifts. Proceedings of the international conference on system science, 36, Hawaii
26. Martins R, Carvalho J (2015) Eye blinking as an indicator of fatigue and mental load – a systematic review. In: Arezes P et al (eds) Occupational safety and hygiene III. Taylor & Francis, London, pp 237–259
27. Matthews G, Campbell S (1998) Task-induced stress and individual differences in coping. Proceedings of the human factors and ergonomics society 42nd annual meeting
28. Misiti M, Misiti Y, Oppenheim G, Poggi J (1996) Wavelet toolbox user's guide. The MathWorks Inc., Natick
29. Parasuraman R, Rizzo M (2007) Neuroergonomics: the brain at work. University Press, Oxford
30. Robert G, Hockey J (1997) Compensatory control in the regulation of human performance under stress and high workload: a cognitive – energetical framework. Biol Psychol 45:73–93
31. Rubio S, Díaz E, Martín J, Puente J (2004) Evaluation of subjective mental workload: a comparison of SWAT, NASA-TLX, and workload profile methods. Appl Psychol Int Rev 53:61–86
32. Sardy S (2000) Minimax threshold for denoising complex signals with waveshrink. IEEE Trans Signal Process 48:102
33. Schleicher R, Galley S, Briest S, Galley L (2008) Blinks and saccades as indicators of fatigue in sleepiness warners: looking tired? Ergonomics 3:2–29
34. Schubert P, Kirchner M (2014) Ellipse area calculations and their applicability in posturography. Gait Posture 39:518–522
35. Schwalm M (2009) Pupillometrie als Methode zur Erfassung mentaler Beanspruchungen im automotive Kontext. Doctoral dissertation, University of Saarland, Saarbrücken
36. Steinhauer S, Siegle G, Condray R, Pless M (2004) Sympathetic and parasympathetic innervation of pupillary dilation during sustained processing. Int J Psychophysiol 52:77–86
37. Straeter O (2005) Cognition and safety: an integrated approach to systems design and assessment. Ashgate, Brussels

38. Watling C, Smith S (2013) In the blink of an eye: the circadian effects on ocular and subjective indices of driver sleepiness. In: Australasian road safety research, policing & education conference, 28–30. Brisbane, QLD
39. Wickens C (2002) Multiple resources and performance prediction. Theor Issues Ergon Sci 3 (2):159–177
40. Wickens C, Hollands J, Parasuraman R, Banbury S (2012) Engineering psychology & human performance. Pearson, München
41. Wilson G (2002) An analysis of mental workload in pilots during flight using multiple psychophysiological measures. Int J Aviat Psychol 12(1):3–18